Repositorios de publicaciones digitales de libre acceso en Europa: análisis y valoración de la accesibilidad, posicionamiento web y calidad del código

Por **Cristòfol Rovira, Mari-Carmen Marcos** y **Lluís Codina**

Cristòfol Rovira es doctor y profesor de la Universidad Pompeu Fabra desde el año 1992 en el área de biblioteconomía y documentación. Imparte docencia en las titulaciones de publicidad y relaciones públicas y de comunicación audiovisual, así como en el Master en Documentación Digital (Universidad Pompeu Fabra) y el Master en Gestión de Contenidos Digitales (Universidad de Barcelona/Universidad Pompeu Fabra). Actualmente investiga en el desarrollo de nuevas herramientas para la gestión del conocimiento, como los mapas conceptuales, y para la evaluación de sedes web. Forma parte del grupo de investigación DigiDoc del Instituto Universitario de Lingüística Aplicada (Universidad Pompeu Fabra). Es codirector del anuario hipertext.net y dirige el Master Online en Documentación Digital.
http://www.mapasconceptuales.com
http://www.observaweb.com
http://www.documentaciondigital.org

Mari-Carmen Marcos es doctora y profesora en la Universidad Pompeu Fabra de Barcelona desde 2002. Coordina e imparte docencia en el Master Online en Documentación Digital, es consultora en los estudios de documentación de la Universitat Oberta de Catalunya, y coordinadora de las revistas El profesional de la información, el Anuario Hipertext.net y el Anuario ThinkEPI. Su primer libro, Interacción en interfaces de recuperación de información (2004), marca la línea de su investigación sobre interfaces y usabilidad en sistemas de información, temas sobre los que ha publicado varios artículos. Colabora con distintas empresas e instituciones públicas en proyectos de mejora de la experiencia de usuario y de optimización de sitios web en buscadores.

Lluís Codina es doctor en ciencias de la información y profesor titular de ciencias de la documentación de la Universidad Pompeu Fabra de Barcelona. Imparte las asignaturas de documentación periodística y documentación en los medios en las facultades de Periodismo y de Comunicación Audiovisual de la UPF. Imparte asignaturas de doctorado en la Universidad Pompeu Fabra, la Universidad de Barcelona y la Universidad del País Vasco en temas de su especialidad: sistemas de información documental, investigación en línea y documentación digital. Coordina el grupo de investigación en documentación digital del Instituto Universitario de Lingüística Aplicada de la UPF.

Resumen: *Se presenta el resultado de un estudio de los sitios web de 230 repositorios de publicaciones digitales de libre acceso de 16 países europeos. Se analiza hasta qué punto cumplen con algunas de las características que podrían hacer de ellos sitios web de calidad más allá de los aspectos de contenido, cuya calidad se da por supuesto o, al menos, no es objeto de análisis en esta ocasión. El objetivo es comprobar si los sitios web directamente vinculados con el compromiso del libre acceso a la información satisfacen criterios de accesibilidad, posicionamiento y calidad del código web, es decir, en este último caso observación de los estándares abiertos. Por tanto, se estudia una serie de elementos que afectan directamente a la accesibilidad, el posicionamiento web y la calidad del código fuente. Entre los resultados obtenidos cabe destacar que la accesibilidad debe mejorarse, el código html solamente es aceptable y el posicionamiento es bueno. Es decir, se observan malas prácticas que no deberían darse en sitios cuyo objetivo es facilitar, precisamente, el libre acceso a la información.*

Palabras clave: *Repositorios digitales, Libre acceso, Accesibilidad, Posicionamiento web, Estándares web.*

Title: Digital repositories of open access publications in Europe: analysis and assessment of accessibility, web positioning and source code quality

Abstract: *This work presents the results of a study of web sites of 230 OA digital repositories in 16 European countries. The analysis addresses the extent to which these repositories fulfil display characteristics of quality sites, over and above the quality of their content. The objective of this study is to verify whether web sites specifically committed to free access to information satisfy the quality criteria of accessibility, web positioning and quality of the source code. The results obtained find that accessibility must be improved upon, source code quality is merely acceptable, and web positioning is good. In short, bad practices were observed that should not occur in sites whose objective is to facilitate the free access to information.*

Keywords: *Digital repositories, Open access, Accessibility, Web positioning, Web standards.*

Rovira, Cristòfol; **Marcos, Mari-Carmen**; **Codina, Lluís**. *"Repositorios de publicaciones digitales de libre acceso en Europa: análisis y valoración de la accesibilidad, posicionamiento web y calidad del código". En: El profesional de la información, 2007, enero-febero, v. 16, n. 1, pp. 24-38.*

Artículo recibido el 30-05-06
Aceptación definitiva: 08-11-06

Este trabajo forma parte del proyecto de investigación "Web semántica y sistemas de información" financiado por el *Mº de Educación y Ciencia* (HUM2004-03162/FILO)

Introducción

Uno de los retos a gran escala que tienen las instituciones de investigación es el de construir repositorios digitales que recojan la información producida por sus miembros. El objetivo final es doble: por un lado aportan una política de conservación de documentos digitales, que por su propio formato podrían tener una existencia efímera en los sitios web donde se encuentran disponibles. Y por otro, se trata de conseguir que los resultados de investigación sean más visibles para la comunidad académica, pues al estar en el repositorio se facilitará su localización y será más probable que sean citados por colegas (**Keefer**, 2005a; **Swan**; **Brown**, 2004).

El modelo elegido para gestionar estos fondos es primordialmente el autoarchivo por parte de los autores, de una forma normalizada siguiendo el protocolo *OAI-PMH* (*Open archive initiative-protocol metadata harvesting*) (**Keefer**, 2005b), combinada con la publicación en revistas revisadas por pares (*peer-reviewed*). Los autores pueden enviar al repositorio documentos en versiones previas a la publicación en revistas (*preprints*) o versiones ya publicadas (*post-prints*), así como otros trabajos que no tengan como finalidad su publicación por los canales habituales, por ejemplo material docente. Los documentos publicados en los repositorios pueden hacerlo bajo licencias de uso libre como las *Creative Commons* (CC), donde los autores deciden el modelo: desde el más tradicional, que es el derecho a uso citando la fuente, hasta la libertad de crear obras derivadas manteniendo la autoría original y sin hacer un uso comercial, siempre que sea bajo CC.

http://creativecommons.org/

El crecimiento del número de repositorios ha llevado a la creación de directorios como *Roar* (*Registry of Open Access Repositories*), de la *Universidad de Southampton* (Reino Unido) y *OpenDOAR*, de las universidades de *Nottingham* (Reino Unido) y *Lund* (Suecia). Estas plataformas describen qué materias cubre cada repositorio, si el material está revisado por pares, su política de preservación, etc. *Roar* es por el momento el más exhaustivo: recoge, a fecha de febrero de 2006, 635 repositorios, de los cuales 259 pertenecen a países de la Unión Europea. Están clasificados en diversos tipos (tabla 1) y aunque la mayoría han sido creados por instituciones de investigación también recoge depósitos de tesis, revistas electrónicas y bases de datos.

http://archives.eprints.org/eprints.php

http://www.opendoar.org/

Este trabajo presenta el resultado de un estudio de los sitios web de 230 repositorios de publicaciones digitales en libre acceso de los 16 países de la Unión Europea que constan en el directorio *Roar* (tabla 2). El objetivo es detectar hasta qué punto cumplen con algunas de las características que actualmente son considerados como criterios habituales de calidad. En concreto, se estudian aspectos que afectan directamente a la accesibilidad, el posicionamiento web y la calidad del código fuente, entendida esta última como la adopción de estándares web para el marcado y la codificación de la página.

Hemos considerado de interés tomar como objeto de estudio este tipo de recursos de información porque son una fuente susceptible de ser de gran utilidad para académicos e investigadores. El hecho de que se encuentren accesibles de forma gratuita para toda la comunidad los hace más interesantes aún si cabe. Por este motivo hemos podido centrarnos en aspectos que determinan la calidad de un sitio y que son distintos (e independientes) de la calidad de los contenidos del mismo. Por ejemplo, se puede contar con un contenido de enorme calidad y sin embargo no ser accesible o estar mal posicionado.

Forma parte del objetivo, por tanto, comprobar hasta qué punto los sitios web directamente vinculados con el compromiso del libre acceso a la información satisfacen criterios de accesibilidad, posicionamiento y calidad del código web que deberían estar estrechamente vinculados con los anteriores, al menos cuando hablamos de la web.

Tipo	Registrados	Porcentaje
Base de datos	4	1,5%
Demostración	7	2,7%
Publicación electrónica	20	7,7%
Tesis	27	10,4%
Investigación	165	63,7%
Otros	36	13,9%
Total	259	100,0%

Tabla 1. Repositorios de países de la Unión Europea por tipo de documentos

Metodología de investigación

1. Recogida de datos

Los datos de este estudio se han obtenido usando *DigiDocSpider* (*DDS*), un programa informático de tipo rastreador de desarrollo propio. Las páginas de un sitio web son analizadas a partir de la url de la página de inicio para extraer por medio de expresiones regulares aquellos elementos del código html que sean de interés. *DDS* también es capaz de enviar la página que esté analizando a los distintos servicios de validación disponibles en internet (*xhtml*, *css*, accesibilidad, etc.), recoger el resultado de estos análisis externos e incorporarlo en su base de datos. Se recopilan de forma automática más de 100 indicadores para cada una de las páginas relativos a tres aspectos: accesibilidad, posicionamiento en los buscadores (causas y resultados) y calidad del código *xhtml*. Las tablas detalladas de todos los aspectos analizados se pueden consultar en la url indicada a continuación.

http://www.observaweb.com/repositorios.htm

Como se adelantaba en la introducción, se han tenido en cuenta todos los sitios web de los depósitos digitales de los países de la Unión Europea registrados en *Roar*. De entre los 259 sitios web registrados, 29 de ellos (11,2%) no han podido ser analizados debido a errores en el servidor, por no permitir ser analizados por un rastreador o por disponer de código html erróneo que ha impedido la labor del rastreador, entre otros motivos (tabla 2).

2. Análisis de datos

Se han realizado diferentes tipos de análisis: por un lado se ha estudiado cada repositorio de forma independiente para determinar su calidad en relación a los tres aspectos mencionados para, a continuación, examinar los datos globales, primero por países y luego por tipos de repositorios.

"Es imprescindible disponer de datos independientes de la página inicial, en especial para el posicionamiento, pero también para accesibilidad y calidad del código dada su especial importancia"

La unidad de análisis con *DDS* de esta investigación es la página principal de cada repositorio. Es imprescindible disponer de datos independientes de la página inicial, en especial para el posicionamiento, pero también para accesibilidad y calidad del código dada su especial importancia. No obstante, está previsto realizar dos trabajos futuros: uno que contemple como

País	Sigla	Registrados	Errores	Analizados
Alemania	de	57	4	53
Austria	at	3	0	3
Bélgica	be	9	2	7
Dinamarca	dk	7	0	7
Eslovenia	si	1	0	1
España	es	12	1	11
Finlandia	fi	4	1	3
Francia	fr	26	3	23
Grecia	gr	2	0	2
Holanda	nl	15	5	10
Hungría	hu	4	0	4
Irlanda	ie	2	0	2
Italia	it	21	4	17
Portugal	pt	4	2	2
Reino Unido	gb	68	6	62
Suecia	se	24	1	23
Totales		259	29	230
Porcentajes		100%	11,2%	88,8%

Tabla 2. Número de repositorios por países y porcentaje de sitios rastreados

unidad de análisis los sitios web completos y otro que considere las páginas web que describen y dan acceso a los documentos almacenados en el repositorio con el fin de analizar también sus etiquetas de metadatos.

3. Obtención de los resultados

Se han conseguido dos tipos básicos de resultados: por un lado un ranking de cada una de las características analizadas y, por otro, resultados globales por países.

Para elaborar cada ranking se han analizado las páginas de los repositorios para extraer los valores relativos a un conjunto de indicadores, como por ejemplo el número de etiquetas *title* vacías o la cantidad de enlaces que apuntan a la página analizada (ver tablas 3, 4, 5, 7, 8, 10, 11, 13 y 14). Las puntuaciones obtenidas de cada indicador han sido convertidas en un valor entre 0 y 10 con el fin de poder sumar más fácilmente los valores de diversos indicadores. También se ha aplicado un peso porcentual a cada indicador para expresar su importancia relativa en el contexto de cada ranking. La puntuación global que permite hacer los rankings es la suma del valor de cada indicador después de ser normalizado y de aplicar el peso correspondiente.

Con respecto a los resultados globales por países, es necesario indicar que no sabemos exactamente cuántos depósitos digitales institucionales existen en los países de la Unión Europea que no estén registrados en *Roar*, por tanto no podemos hacer estimaciones sobre la representatividad de los datos analizados en relación a todos los repositorios de un determinado país, aunque estimamos que el número debe ser muy bajo. En todo caso, los valores de los países con un gran número de sitios web analizados son relativamente más fiables para derivar conclusiones sobre el global de los repositorios de cada país. Por la misma razón, los resultados relativos a países con pocos sitios web identificados en *Roar* (Dinamarca, Portugal, Hungría, Finlandia, Austria, Grecia y Eslovenia) deberán tomarse con cautela en el momento de extraer conclusiones generales.

Resultados obtenidos en cuanto a la accesibilidad

La accesibilidad es un elemento básico que los sitios web deberían proponerse cumplir. Se trata de que las páginas puedan ser leídas por cualquier persona independientemente de sus circunstancias personales (discapacidades físicas o sensoriales) y tecnológicas (hardware y software que utilice). Existen herramientas que revisan de forma automática algunas características del código fuente que ponen de manifiesto problemas que afectan a la accesibilidad, como *Hera*, *TAW* o *Wave*, entre otras.

http://www.sidar.org/hera/

http://www.tawdis.net/taw3/cms/es

http://www.wave.webaim.org/wave/

El *W3C* establece tres niveles prioridad en relación a la accesibilidad en páginas web en la *Iniciativa de accesibilidad en la web*, con cada uno de ellos abarcando un mayor nivel de colectivos:

http://www.w3.org/WAI/

— Prioridad 1: son accesibles para buena parte de los colectivos de discapacitados, pero algunas personas con algún tipo de discapacidad pueden tener problemas de acceso.

— Prioridad 2: hay más accesibilidad que en el caso anterior, pero aún puede haber colectivos que no puedan acceder los sitios.

— Prioridad 3: en teoría, son sitios totalmente accesibles, o al menos, agotan las posibilidades de la tecnología actual en este sentido, en principio ningún colectivo queda excluido de ellos.

En su web, el *W3C* establece con total precisión cuáles son los componentes de cada nivel. Queda fuera del alcance y objetivos de este trabajo entrar en ellos, pero para dar una idea de los mismos señalemos que la necesidad de utilizar el atributo *alt* en todas las imágenes y animaciones de un sitio, así como añadir transcripciones de audio y descripciones de vídeo, son algunos de los requerimientos más significativos de la accesibilidad.

El cumplimiento de cada una de estas prioridades implica niveles de accesibilidad correspondientes (cumplir la prioridad 1 implica un sitio con nivel de accesibilidad 1 y así sucesivamente). Para este estudio hemos tomado los resultados que dan los tests *Hera*, *TAW* y *Wave*. A continuación se indican los resultados (gráfico 1 y tablas 3-4). Los países están ordenados en función de la cantidad de repositorios analizados y de forma descendiente.

La presencia de errores de nivel 1 de accesibilidad en los sitios web de los repositorios analizados es relativamente aceptable si usamos el analizador *HERA*. Con él, en promedio global se han localizado tan sólo 0,75 errores por página. En cambio, este dato se dispara hasta los 2,87 si usamos *TAW*.

Los criterios de ambos no son coincidentes, se encuentran casi 4 veces más errores de nivel 1 en *TAW* que en *HERA*. Esta discrepancia se acrecienta con los de nivel 2, donde *HERA* detecta 6,03 y *TAW* 40,07. Esta última cifra resulta sorprendentemente alta. Desconocemos los motivos concretos de estas diferencias que, obviamente, derivan de la aplicación de distintas formas de medición, a pesar de que ambos servicios dicen ajustarse a las recomendaciones del *W3C*.

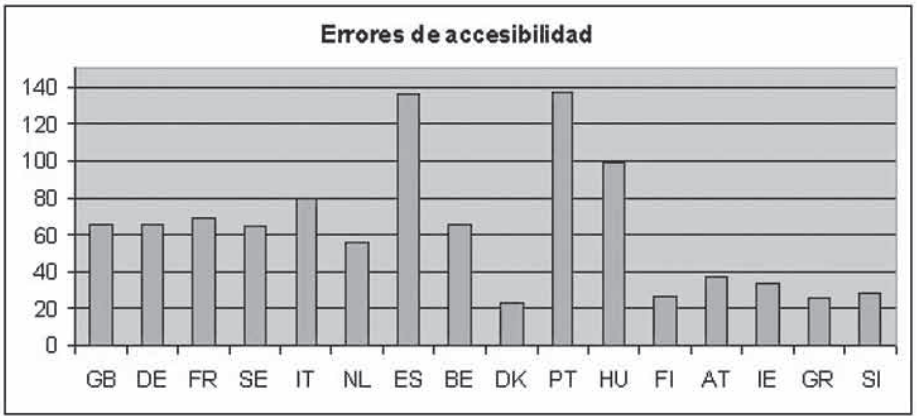

Gráfico 1. Errores de accesibilidad por países

	Indicadores de accesibilidad	gb	de	fr	se	it	nl	es	be
1	Errores (*Hera* 1)	0,40	0,96	1,13	0,48	0,88	1,00	1,18	1,14
2	Errores (*Hera* 2)	5,87	5,26	6,17	5,13	6,00	5,40	6,64	5,86
3	Errores (*Hera* 3)	0,40	0,96	1,13	0,48	0,88	1,00	1,18	1,14
4	Errores (*TAW* 1)	2,29	8,43	3,22	2,00	2,65	3,80	7,36	5,43
5	Errores (*TAW* 2)	43,32	34,15	46,87	37,91	52,53	34,10	96,91	35,43
6	Errores (*TAW* 3)	8,94	6,68	6,57	10,17	12,65	6,90	13,27	10,57
7	Errores (*WAVE*)	3,77	8,74	4,22	7,70	4,29	3,60	9,00	6,14
8	Suma de errores	64,99	65,18	69,31	63,87	79,88	55,8	135,54	65,71
9	**Valor global**	**4,13**	**3,59**	**4,00**	**4,66**	**3,71**	**4,10**	**3,41**	**3,09**

Tabla 3. Análisis de la accesibilidad por países (1)

	Indicadores de accesibilidad	dk	pt	hu	fi	at	ie	gr	si	Pro-medio global	Total global
1	Errores (*Hera* 1)	0,57	2,00	1,25	0,00	1,00	0,00	0,00	0,00	0,75	11,99
2	Errores (*Hera* 2)	5,29	7,50	7,25	5,00	4,67	7,00	6,50	7,00	6,03	96,54
3	Errores (*Hera* 3)	0,57	2,00	1,25	0,00	1,00	0,00	0,00	0,00	0,75	11,99
4	Errores (*TAW* 1)	2,14	3,00	2,25	0,00	3,33	0,00	0,00	0,00	2,87	45,90
5	Errores (*TAW* 2)	9,57	98,50	74,75	14,33	19,67	18,50	11,50	13,00	40,07	641,04
6	Errores (*TAW* 3)	4,29	17,50	10,75	5,00	7,00	6,50	2,50	6,00	8,46	135,29
7	Errores (*WAVE*)	1,00	7,00	1,75	2,33	1,00	1,00	5,50	2,00	4,32	69,04
8	Suma de errores	23,43	137,50	99,25	26,66	37,67	33,00	26,00	28,00	63,24	1011,79
9	**Valor global**	**4,81**	**1,72**	**3,50**	**5,04**	**4,03**	**4,12**	**4,55**	**4,21**	**3,95**	

Tabla 4. Análisis de la accesibilidad por países (2)

Los países con un mayor número de errores de accesibilidad son Portugal, España, Hungría e Italia. Si consideramos tan sólo los errores de nivel 1, los peores países son Alemania, España, Bélgica y Austria. La situación de los sitios web de los depósitos digitales españoles es especialmente crítica en accesibilidad con el doble de errores, tanto en número de global como de nivel 1.

"La situación de los sitios web de los depósitos digitales españoles es especialmente crítica en accesibilidad con el doble de errores"

Los principales motivos de que sea así están relacionados con los fallos de codificación html, con el uso inadecuado (o no uso) de *css*, la presencia de etiquetas desaconsejadas por el *W3C* y la codificación incompleta o errónea de las etiquetas de las imágenes, formularios, enlaces y tablas (tabla 5).

Los valores de las tablas 3, 4 y 5 son una selección de entre todos los indicadores de accesibilidad obtenidos. A cada indicador se le asignado un peso porcentual para calcular un valor global entre 0 y 10 que expresa la calidad de la accesibilidad de cada repositorio. Este valor local ha permitido obtener un valor global de cada país en relación a la accesibilidad (fila 9 de las tablas 2 y 3) y por otro ha permitido establecer un ranking general de todos los repositorios con independencia del país al que pertenecen.

La puntuación global de la accesibilidad está por debajo de 5 sobre 10 en todos los países exceptuando Finlandia. El promedio global es de tan sólo de 3,95 sobre 10 (fila 9 de las tablas 3 y 4). La conclusión es que es necesario revisar a fondo este aspecto de la implementación de los sitios web de los repositorios para poder ofrecer un servicio adecuado a los usuarios, ya sea los que tienen necesidades especiales como los que no.

En el ranking de accesibilidad aparece uno de los pocos repositorios españoles presentes en los rankings elaborados. Se trata de *SORT*, un repositorio de tipo publicación electrónica publicado por el *Instituto de Estadística de Cataluña* que también está presente en el ranking sobre la calidad del código *xhtml*.

Posicionamiento

1. Presencia en buscadores y directorios

El posicionamiento es uno de los aspectos que más debería preocupar a los responsables de sitios web. Se trata de conseguir una buena posición en la página de resultados de los buscadores cuando un usuario realiza una consulta sobre la que el sitio web en cuestión pue-

Indicadores de error en accesibilidad	Valor global obtenido	Valor recomendado
Presencia de etiquetas desaconsejadas por *W3C*	25,3 *	0
Ausencia de etiquetas *doctype*	22,6%	100%
Presencia de errores html según el validador de *W3C*	19,5 *	0
Presencia de errores hojas de estilo según el validador de *W3C*	3,5 *	0
Ausencia de etiquetas *h1*	0,6 *	3
Ausencia de etiquetas *h2*	0,4 *	2
Ausencia de etiquetas *h3*	0,5 *	2
Ausencia de etiquetas *h4*	0,1 *	1
Ausencia de etiquetas *img* con atributo *alt*	47,0%	100%
Ausencia de etiquetas *img* con atributos *width*	67,6%	100%
Ausencia de etiquetas *input* de formulario con *acceskey*	0,0%	100%
Ausencia de etiquetas *input* de formulario con *tabindex*	1,6%	100%
Presencia de etiquetas de enlace con atributo *target _blank*	4,9%	0%
Ausencia de etiquetas de enlace con atributo *accesskey*	1,0%	100%
Ausencia de etiquetas de enlace con atributo *alt*	0,1%	100%
Presencia de etiquetas de enlace con *javascript*	0,5%	0%
Ausencia de etiquetas de enlace con atributo *title*	14,6%	100%
Ausencia de etiquetas *noscript*	6,9%	100%
Ausencia de etiquetas de tablas con parámetro *caption*	0,0%	100%
Ausencia de etiquetas de tablas con atributo *summary*	3,3%	100%
Ausencia de etiquetas de tablas con etiquetas *th*	1,8%	100%
Ausencia de etiquetas de tablas con etiquetas *title*	0,0%	100%
* Promedio por página de los valores del indicador		

Tabla 5. Principales motivos de los errores de accesibilidad

	Accesibilidad: los mejores 10 sitios web de repositorios	País	Puntuación
1	*Academic Archive On-line* (*Aarhus University*, Denmark)	dk	6,8898
2	*Statistics and Operations Research Transactions–SORT*	es	6,8504
3	*Academic Archive On-line* (*Umea University*, Sweden)	se	6,8307
4	E-ms Eprints Open Archive in Social Medicine and related fields	*it*	6,2992
5	Cambridge University Computer Science Technical Reports	*gb*	5,9449
6	*Academic Archive On-line* (*Uppsala University*, Sweden)	se	5,9252
7	École Normale Supérieure: Lettres et sciences humaines (Human Sciences Archive)	fr	5,8661
8	XIOS Hogeschool Limburg	*be*	5,8465
9	Swedish Institute of Computer Science Publications Database	se	5,8268
10	Lund University Publications	*se*	5,8268

Tabla 6. Ranking de repositorios en función de la accesibilidad

de ofrecer información de utilidad. El posicionamiento web de tipo "ético" se consigue con algunas indicaciones que hemos tenido en cuenta en este análisis siguiendo el trabajo de **Lluís Codina** y **Mari-Carmen Marcos** (2005).

Uno de los criterios de peso es la popularidad del sitio en cuestión, en concreto el número de citas (enlaces) que recibe de otras páginas web. En este sentido, *Google* ha creado un indicador denominado *PageRank* (*PR*); se trata de una puntuación que oscila entre 0 y 10 y que indica el grado de popularidad que tiene una página web. Para obtenerlo tiene en cuenta tanto el número de citas recibidas como el *PR* que tiene asignado a su vez cada una de las páginas citantes. Un aspecto relacionado con éste es la aparición de los sitios web en grandes directorios como *Dmoz* o *Yahoo!*, que tienen a su vez un alto *PR*: si un sitio web aparece recogido en estos directorios, su propio *PR* tenderá a aumentar. Igualmente, si no es el repositorio el que se encuentra en estos directorios sino la institución que lo gestiona, también será éste un rasgo positivo, ya que heredará en parte el *PR* conseguido por su institución. Por lo tanto, y en cuanto al posicionamiento, en este estudio hemos tomado estos indicadores:

1. *Dmoz* dominio: porcentaje de dominios que están en el directorio *Dmoz*.

2. *Yahoo!* (di) dominio: porcentaje de dominios que están el directorio *Yahoo!*.

3. *Google* dominio: promedio de páginas del dominio indexadas en *Google*.

4. *Yahoo!* (bu) dominio: promedio de páginas del dominio indexadas en *Yahoo!* (buscador).

5. *Dmoz* página: porcentaje de páginas analizadas que están en el directorio *Dmoz*.

6. *Yahoo!* (di) página: porcentaje de páginas analizadas que están en el directorio *Yahoo!*.

Por el mismo motivo, otro dato de interés para evaluar el posicionamiento de un sitio es el número de enlaces en entrada (o citas) que recibe, así como el número de enlaces de salida (hacia otras páginas). Para medirlo hemos considerado:

7. Links entran G: promedio por página del número de enlaces entrantes (emitidos desde otras páginas hacia la página analizada) según *Google*.

8. Links entran Y: promedio por página del número de enlaces entrantes según *Yahoo!*.

9. Links entran Y (ext): promedio por página del número de enlaces entrantes según *Yahoo!* de páginas externas al dominio. Este indicador corresponde a la visibilidad.

10. Link salen (ext): promedio por página del número de enlaces salientes externos (enlaces desde la página web analizada hacia otros sitios web).

11. Link salen (Int): promedio por página del número de enlaces salientes internos (enlaces desde la página web analizada hacia otras páginas del mismo sitio web).

12. Luminosidad: total de enlaces salientes.

Así mismo, teniendo en cuenta que algunas características del código fuente facilitan (o dificultan) la indexación por parte de buscadores, se ha considerado de interés analizar si estos repositorios cuentan con etiquetas descriptivas de contenido, como el título de la página (*title*), el texto alternativo en las imágenes y en los enlaces (*alt*) y el título de los enlaces (*title*):

13. *Title* vacías: porcentaje de páginas con etiquetas *title* vacías.

14. Gráfico *alt*: porcentaje de gráficos con atributo *alt.*

15. Links *alt*: porcentaje de enlaces con atributo *alt.*

16. Links *title*: porcentaje de enlaces con atributo *title*.

Por último, aunque no es una barrera definitiva, se ha considerado que el hecho de contar con frames dificulta en parte la indexación correcta por parte de los motores, por lo tanto se ha recogido también este dato:

17. Frames: promedio por página del número de frames.

Las tablas 7 y 8 muestran los resultados obtenidos para cada país en cuanto a estos criterios que se han comentado. Para calcular la puntuación global se ha realizado el promedio a partir de todos los indicadores obtenidos y aplicando un peso porcentual a cada uno de ellos. Se pueden consultar los pesos asignados en la web indicada a continuación.

http://www.observaweb.com/repositorios.htm

De entre los indicadores sobre las causas del posicionamiento hay que destacar la baja presencia de los repositorios analizados en los directorios *Dmoz* y *Yahoo!*, tanto de las url de los dominios donde están hospedados (universidades, instituciones de investigación, publicaciones digitales, etc.) como de las direcciones de las páginas iniciales de cada repositorio (filas 1, 2, 5 y 6 de las tablas 7 y 8). En cualquiera de los casos no se alcanza una cifra superior al 24%. La presencia en *Yahoo!* es siempre tres veces inferior que en *Dmoz*.

Otro aspecto que se debe mejorar es la presencia de los atributos *alt* y *title* en gráficos y enlaces. El 66% de los gráficos no tienen el atributo *alt* y el 91% de los enlaces no llevan *title*. Además de mejorar la accesibilidad, son usados por los buscadores para aumentar la pertinencia de una página cuando las palabras de las búsquedas coinciden con su texto. Por tanto, es altamente recomendable su presencia para mejorar el posicionamiento.

Como aspectos positivos hay que resaltar la gran cantidad de url de los dominios que están presentes en los buscadores, con promedios de 50.782 y 17.616 respectivamente. Estos datos muestran que, en general, los repositorios forman parte de enormes sitios web, como es el caso de universidades o institutos de investigación. Este factor favorece la conectividad de las páginas de inicio de los repositorios con gran cantidad de enlaces entrantes. Como veremos en el siguiente apartado, los valores de *PR* obtenidos son altos. Una de las principales razones es precisamente la gran cantidad de enlaces entrantes externos, con una media 422 según *Yahoo!*.

	Indicadores de las causas de posicionamiento	**gb**	**de**	**fr**	**se**	**it**	**nl**	**es**	**be**
1	*Dmoz* dominio	18%	30%	48%	61%	18%	30%	55%	0%
2	*Yahoo!* (di) dominio	6%	4%	9%	4%	0%	0%	9%	0%
3	*Google* dominio	37.432,53	51.390,42	133.958,70	42.299,13	23.849,24	111.387,00	327.267,45	20.100,86
4	*Yahoo!* (bu) dominio	7.576,79	7.870,23	30.387,22	4.336,13	2.579,29	16.182,90	174.031,18	966,86
5	*Dmoz* página	16%	19%	43%	9%	18%	10%	18%	0%
6	*Yahoo!* (di) página	5%	4%	4%	4%	0%	0%	0%	0%
7	Links entran G	610,98	165,36	832,09	132,57	220,94	402,70	896,91	74,29
8	Links entran Y	5.448,97	1.277,94	4.991,52	1.030,39	1.350,12	3.083,50	1.316,18	73,57
9	Links entran Y (ext)	439,58	207,17	2.718,61	46,00	222,88	418,20	183,18	55,00
10	Link salen (ext)	8,69	6,89	8,09	7,65	5,59	4,70	5,64	3,57
11	Link salen (int)	18,95	14,45	22,09	6,13	17,12	14,40	22,91	18,86
12	*Title* vacías	0%	4%	4%	0%	0%	0%	0%	0%
13	Gráfico *alt*	58%	40%	44%	60%	38%	16%	52%	20%
14	Links *alt*	0%	0%	0%	0%	0%	0%	0%	0%
15	Links *title*	3%	4%	3%	19%	1%	3%	2%	0%
16	Frames	0,00	0,26	0,22	0,00	0,00	0,70	0,82	0,00
17	Luminosidad	27,65	21,34	30,17	13,78	22,71	19,10	28,55	22,43
18	**Puntuación global**	**6,02**	**6,21**	**6,31**	**6,10**	**6,32**	**5,74**	**6,34**	**5,61**

Tabla 7. Análisis de las causas del posicionamiento (1)

	Indicadores de las causas de posicionamiento	dk	pt	hu	fi	at	ie	gr	si	Promedio global
1	*Dmoz* dominio	14%	0%	0%	0%	67%	50%	0%	0%	24%
2	*Yahoo!* (di) dominio	0%	0%	25%	33%	33%	0%	0%	0%	8%
3	*Google* dominio	20516,71	17046,00	83,75	371,33	12340,33	364,00	13700,00	408,00	50782,22
4	*Yahoo!* (bu) dominio	3024,43	122,50	559,50	1238,33	1953,33	587,00	30000,00	451,00	17616,67
5	*Dmoz* página	14%	0%	0%	0%	0%	50%	0%	0%	19%
6	*Yahoo!* (di) página	0%	0%	25%	0%	33%	0%	0%	0%	5%
7	Links entran G	567,00	17,50	9,75	33,33	88,00	60,00	88,50	45,00	265,31
8	Links entran Y	1431,43	135,00	93,75	69,67	1184,33	452,50	166,00	271,00	1398,49
9	Links entran Y (ext)	1380,57	76,00	32,50	28,67	756,33	165,00	11,50	22,00	422,70
10	Link salen (ext)	3,43	2,50	3,75	7,67	3,33	3,00	1,50	6,00	5,12
11	Link salen (Int)	12,71	48,50	18,25	12,67	7,67	15,50	8,00	25,00	17,70
12	*Title* vacías	0%	0%	0%	0%	0%	0%	0%	0%	1%
13	Gráfico *alt*	50%	0%	60%	30%	42%	0%	0%	0%	44%
14	Links *alt*	0%	0%	0%	0%	0%	0%	0%	0%	0%
15	Links *title*	13%	0%	0%	0%	0%	0%	0%	0%	9%
16	Frames	2,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,31
17	Luminosidad	16,14	51,00	22,00	20,33	11,00	18,50	9,50	31,00	16,14
18	Puntuación global	5,21	6,42	4,82	5,35	6,91	8,10	7,07	8,07	**6,29**

Tabla 8. Análisis de las causas del posicionamiento (2)

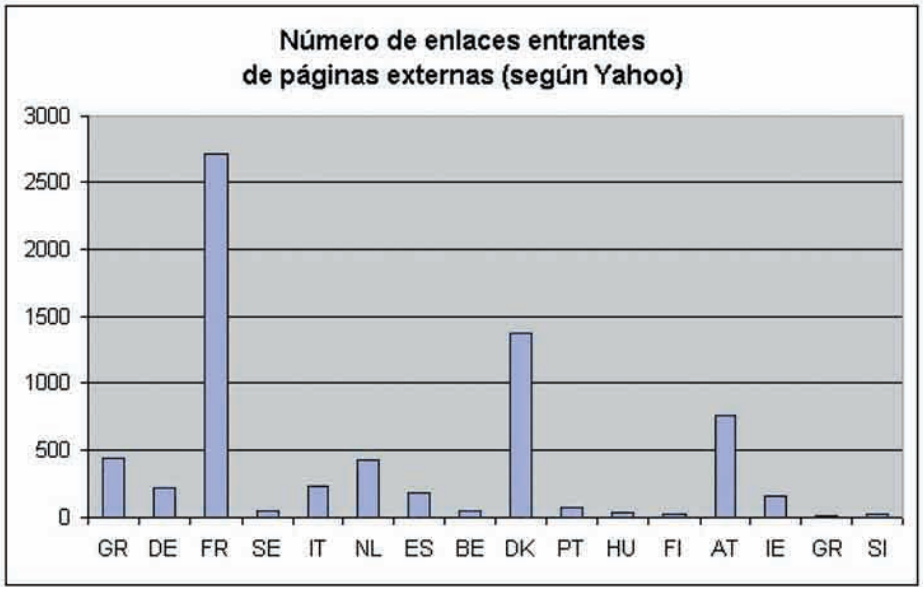

Gráfico 2. Número de enlaces entrantes de páginas externas por países

	Causas posicionamiento: los mejores 10 sitios web de repositorios	País	Puntuación
1	Resource Discovery Network	gb	9,1102
2	Organic Eprints	dk	8,9961
3	Living Reviews in Relativity	de	8,8307
4	Psycoloquy (Journal)	gb	8,7992
5	Digital Peer Publishing	de	8,7165
6	NDAD-UK National Digital Archive of Datasets	gb	8,6969
7	Revues,org-Fédération de revues en ligne en sciences humaines et sociales	fr	8,6024
8	E-Lis: Research in Computing and Library and Information Science	it	8,6024
9	HAL: Hyper Article en Ligne	fr	8,5630
10	E-Prints Universidad Complutense Madrid	es	8,5039

Tabla 9. Ranking de los repositorios en función de las causas del posicionamiento

Finalmente resaltar también como aspectos positivos la casi nula presencial de frames y de etiquetas *title* vacías. Estos elementos entorpecen la tarea de los buscadores disminuyendo el posicionamiento.

En el análisis por países cabe indicar la gran cantidad de enlaces entrantes externos en Francia, con 2.718 cuando la media está en 422. Esta proporción no se mantiene en los resultados de *PR*, por tanto, probablemente el *PR* de las páginas que emiten estos enlaces será bajo.

Hay poca variabilidad en la puntuación global, con valores entre 5,5 y 6,5 en todos los países exceptuando Irlanda, Grecia y Eslovenia. El escaso número de sitios web analizados de estos tres países podría ser la causa de estos valores relativamente altos.

En el ranking de los 10 mejores sitios web en relación a las causas del posicionamiento podemos encontrar a *E-Prints* de la *Universidad Complutense Madrid* en décimo lugar. Este repositorio, junto con *SORT*, son los únicos que aparecen en los rankings elaborados.

2. Indicadores de visibilidad

En el apartado anterior se han presentado los principales factores que de acuerdo a la bibliografía influyen en el posicionamiento de un sitio. Ahora revisaremos los resultados de este posicionamiento usando el *PR* como indicador de visibilidad, es decir, de la facilidad teórica para encontrar un sitio. La mediana de *PR* nos permite valorar de forma global el posicionamiento por países.

En los valores de *PR* podemos constatar una variación relativa. Por un lado, tenemos en general puntuaciones entre 4 y 5,5 en todos los países, excepto Portugal (3) y Austria, Reino Unido, Francia y Suecia (6). Debemos recordar, no obstante, que el *PR* corresponde a una escala logarítmica, por lo cual, en realidad no solamente un *PR* de 3 es muy distinto de uno de 4, sino que dos sitios con el mismo valor pueden tener en realidad un número de enlaces de entrada muy distinto. *Google* no facilita el dato "real", sino únicamente esta escala de tipo logarítmico de va de 0 a 10 ya mencionada (puede verse una discusión sobre el *PR* "real" y el "aparente" en **Codina** y **Marcos**, 2005).

En todo caso, creemos que son valores relativamente altos teniendo en cuenta que a pesar de que la puntuación máxima del *PR* es 10, hay en el mundo pocas sedes web que superen los 7 puntos.

En el ranking por países (tabla 12) hemos seleccionado los sitios web de repositorios con un *PR* igual o superior a 7. Hay un total de 26 casos que cumplen con esta condición y casi un 50% corresponden a repositorios del Reino Unido, debido en gran parte al prestigio de las instituciones que los albergan.

Resultados relacionados con la calidad del código html

Html ha experimentado un salto cualitativo con la versión *xhtml*; el hecho de usar éste por aquel carece de impacto directo en el posicionamiento, pero *xhtml* impone una codificación más rigurosa y lógica, por tanto,

Page Rank	gb	de	fr	se	it	nl	es	be
Page Rank	6	5	6	6	5	5	5	5

Tabla 10. Posicionamiento (1)

Page Rank	dk	pt	hu	fi	at	ie	gr	si
Page Rank	5	3	5	5	6	5,5	5,5	5

Tabla 11. Posicionamiento (2)

	Posicionamiento: los 25 mejores sitios web de repositorios		Page Rank
1	Gallica, bibliothèque numérique de la Bibliothèque nationale de France	fr	8
2	Resource Discovery Network	gb	8
3	Academic Archive On-line (Stockholm University, Sweden)	su	7
4	Academic Archive On-line (Uppsala University, Sweden)	su	7
5	Archive Ouverte Inria	fr	7
6	Behavioral and Brain Sciences (Journal)	gb	7
7	Cclrc ePublication Archive	gb	7
8	CCSD: TEL (doctor self-archived theses)	fr	7
9	CogPrints Cognitive Sciences Eprint Archive	gb	7
10	Cultivate Interactive	gb	7
11	DSpace at Cambridge	gb	7
12	Electronic Resource Preservation and Access Network: ErpaePrints Service	gb	7
13	E-Lis: Research in Computing and Library and Information Science	it	7
14	GSI Gesellschaft für Schwerionenforschung mbH	de	7
15	Living Reviews in Relativity	de	7
16	Numdam-Numérisation de documents anciens mathématiques	fr	7
17	Oxford Eprints	gb	7
18	Persee: Revues scientifique en sciences humaines et sociales	fr	7
19	Psycoloquy (Journal)	gb	7
20	Revues,org-Fédération de revues en ligne en sciences humaines et sociales	fr	7
21	The Arts and Humanities Data Service	gb	7
22	Universiteit van Tilburg	nl	7
23	University College London Eprints	gb	7
24	University of Southampton: Department of Electronics and Computer Science	gb	7
25	University of Trento: Unitn-eprints	it	7
26	VTT Technical Research Centre of Finland	fi	7

Tabla 12. Ranking de los repositorios en función del Page Rank

incrementa su facilidad de procesamiento por parte no solamente de los motores de búsqueda, sino de cualquier software de análisis. Adicionalmente, su uso es un indicador significativo del compromiso del sitio con el cumplimiento de estándares.

En este apartado hemos tenido en cuenta algunos indicadores que ponen de manifiesto la calidad del código de la página principal de consulta de los repositorios, como es la existencia de la declaración *doctype* en la página, el uso de valores de atributos entre comillas, que no contengan etiquetas desaconsejadas (*deprecated*) por el *W3C*, que el código *xhtml* y de las hojas de estilo *css* no tengan errores al pasarlo por los validadores del *W3C*, que las etiquetas están en minúsculas y que no haya enlaces internos rotos. Las tablas 13 y 14 muestran los resultados obtenidos así como la puntuación global, que se ha obtenido aplicando a cada indicador un peso porcentual y calculando el promedio.

El promedio global de errores de las páginas analizadas es muy alto, con promedios de 18,5 por página en html y 6,6 en las hojas de estilo. A esta deficiente codificación hay que añadir un uso generalizado de etiquetas desaconsejadas (*deprecated*) por el *W3C* por ser obsoletas, con un sorprendente promedio de 22,57 por página. También es remarcable la ausencia de comillas en los parámetros con un promedio de 15,03 por página.

Como elementos positivos hay que resaltar la ausencia total de enlaces rotos, la tendencia generalizada (88%) de usar letras minúsculas en las etiquetas y el alto porcentaje de la etiqueta *doctype* con un 70% de los sitios web analizados.

	Indicadores de html (%)	gb	de	fr	se	it	nl	es	be
1	Presencia de *doctype*	87	57	74	87	100	60	82	100
2	Sin comillas	5,37	9,55	23,22	11,35	9,06	14,20	17,09	39,71
3	Presencia de etiquetas desaconsejadas	28,21	22,26	27,48	24,35	25,35	11,60	54,09	20,00
4	Errores html	13,65	23,55	29,04	15,09	15,88	20,10	31,73	38,71
5	Errores *css*	2,95	2,75	1,22	3,61	3,94	1,20	8,36	7,14
6	Minúsculas	98	85	89	97	0,94	82	93	88
7	Enlaces rotos	0	0	0	0	0	0	0	0
8	**Puntuación global**	**5,44**	**4,79**	**5,43**	**6,37**	**5,64**	**6,03**	**4,8**	**5,91**

Tabla 13. Análisis de la calidad del código xhtml *(1) en %*

	Indicadores de html (%)	dk	pt	hu	fi	at	ie	gr	si	Promedio global
1	Presencia de *doctype*	71	50	100	100	100	0	50	0	70
2	Sin comillas	6,29	77,00	1,50	3,00	6,67	0,50	12,00	4,00	15,03
3	Presencia de etiquetas desaconsejadas	8,71	34,50	41,25	8,67	11,67	13,00	21,00	9,00	22,57
4	Errores html	10,71	40,50	8,25	3,33	10,33	17,00	6,50	13,00	18,59
5	Errores *css*	1,86	64,50	1,00	0,67	2,33	4,00	0,00	0,00	6,60
6	Minúsculas	97	76	100	98	62	100	50	100	88
7	Enlaces rotos	0	0	0	0	0	0	0	0	0
8	**Puntuación global**	**6,65**	**3,49**	**6,92**	**7,07**	**7,00**	**5,13**	**6,75**	**6,18**	**5,85**

Tabla 14. Análisis de la calidad del código xhtml *(2) en %*

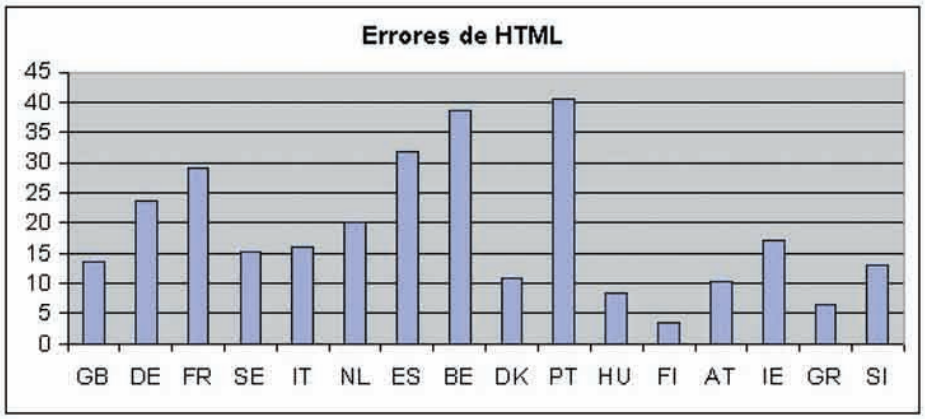

Gráfico 3. Errores de html por países

Los países con un mayor número de errores en el código fuente *xhtml* son Portugal, Bélgica, España y Francia. Los que han obtenido un menor valor global son Alemania, Portugal, España y Francia. No obstante, la puntuación global de la calidad del código es aceptable (5,85). En todos los países (exceptuando Portugal con un 3,49) prácticamente superan el 50% de la puntuación posible y en algunos casos, como Finlandia y Austria se obtienen valores superiores al 70%. En la tabla 15 se puede consultar los 10 mejores sitios web en calidad del código, todos ellos con puntuaciones superiores al 8. En el ranking de los diez mejores sitios web vuelve a aparecer el caso español de *SORT*.

Conclusiones

En cuanto a la accesibilidad y calidad del código fuente, en términos generales los resultados de este estudio ponen de manifiesto que la accesibilidad de los sitios web de repositorios en Europa debe mejorarse

ya que se obtienen valores promedio globales de tan sólo 3,95 sobre 10. Los principales motivos de esta baja puntuación son deficiencias en la codificación *xhtml* y *css*, presencia de etiquetas desaconsejadas por el W3C y la codificación incompleta o errónea de las etiquetas de las imágenes, formularios, enlaces y tablas.

"La accesibilidad de los sitios web de repositorios en Europa debe mejorarse ya que se obtienen valores promedio globales de tan sólo 3,95 sobre 10"

En este sentido, hay un número relativamente alto de errores de *xhtml*, de *css* y uso de etiquetas obsoletas, con valores promedios por página de 18,59, 6,60 y 22,57 respectivamente. La puntuación global en este apartado (5,85 sobre 10) aunque es aceptable requeriría un esfuerzo de mejora. Esto, probablemente, es más indicador de una deficiente comprensión de la accesibilidad que de un desinterés por el tema. Por lo tanto, creemos, que los responsables de esta clase de sitios estarían muy predispuestos a su mejora en este aspecto con una formación adecuada respecto de lo que significa accesibilidad. Aquí, por tanto, la recomendación es clara: conviene hacer un llamamiento para que los responsables académicos y webmaster asuman competencias sobre el tema.

En lo que se refiere al posicionamiento web, los resultados son relativamente altos a pesar de la poca presencia de los repositorios en los directorios *Dmoz* y *Yahoo!*. Probablemente, los altos valores de luminosidad y visibilidad propician este buen posicionamiento. Esto, con seguridad, indica al menos dos cosas: por un lado, un buen contenido y su actualización frecuente son causa de buen posicionamiento en general. En segundo lugar, los responsables de esta clase de sitios suelen estar a su vez bien motivados para cuidar los aspectos vinculados con el posicionamiento. No obstante, parece recomendable que lleven a cabo políticas de posicionamiento más sistemáticas y que, entre otras cosas, conduzcan a su inclusión sistemáticas en directorios de amplio uso, como *Dmoz*, *Yahoo!* y otros.

Trabajo futuro

En lo sucesivo, creemos que la línea de investigación que hemos presentado en este trabajo se puede ampliar y mejorar con el análisis del sitio completo (o de varios niveles completos del sitio) en lugar de limitarlo a la página principal, aunque esta última tenga una importancia singular.

También está previsto realizar análisis comparativos de los repositorios institucionales que utilizan diferente software con el fin de determinar si existe una relación entre éste y los resultados ya obtenidos en el presente estudio.

Así mismo, sería conveniente llevar este tipo de estudios a repositorios de otros países y zonas geográficas; en particular, parece interesante ampliarlo a repositorios de EUA por su especial peso en el mundo de la ciencia, sin perjuicio de considerar otras áreas de gran interés como Asia, Australia y América Latina.

La clase de análisis señalados permitiría estudios comparados entre Europa y otras regiones del mundo. En cualquier caso, nuestra intención es repetir el estudio presentado aquí con cierta periodicidad con el objetivo de disponer en el futuro de investigaciones longitudinales que permitan apreciar las líneas de evolución en el tiempo de los aspectos aquí estudiados. Por último, para no alargar los ejemplos, podemos mencionar el estudio del uso de metadatos específicos para describir los documentos incluidos en los repositorios, así como analizar su encaje en los planes de la futura web semántica que promociona el *W3C*.

Referencias

Codina, Lluís; **Marcos, Mari-Carmen**. "Posicionamiento web: conceptos y herramientas". En: *El profesional de la información*, 2005, v. 14, n. 2, pp. 84-99.

Keefer, Alice. "Los autores y el self-archiving". En: *ThinkEPI*, 2005a. http://www.thinkepi.net/repositorio/los-autores-y-el-self-archiving/

Keefer, Alice. "Aproximació al moviment open access". En: *BiD*, 2005b, n. 15.
http://www2.ub.edu/bid/consulta_articulos.php?fichero=15keefer.htm

Swan, A.; **Brown, S.** (2005). "Open access self-archiving: an author study". Technical report, External Collaborators, Key Perspectives Inc., 2005. http://eprints.ecs.soton.ac.uk/10999/

Van Westrienen, Gerard; **Lynch, Clifford A.** "Academic institutional repositories: deployment status in 13 nations as of mid 2005". En: *D-lib magazine*, 2005, v. 11, n. 9.
http://www.dlib.org/dlib/september05/westrienen/09westrienen.html

World Wide Web Consortium. Oficina Española. Guía breve de accesibilidad web.
http://www.W3C.es/divulgacion/guiasbreves/Accesibilidad

***Cristòfol Rovira**, **Mari-Carmen Marcos** y **Lluís Codina**, Dpto. de Periodismo y Comunicación Audiovisual, Universidad Pompeu Fabra.*
cristofol.rovira@upf.edu
mcarmen.marcos@upf.edu
lluis.codina@upf.edu